Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 14th International Scientific and Practical Conference. Volume 2, 12-17

Internet Traffic Zone Identification by Backpropagation and Probabilistic Neural Networks

Ivelina Balabanova
Department of Communications Equipment and Technologies
Technical University of Gabrovo
Gabrovo, Bulgaria
ivstoeva@abv.bg

Teodora Zhorova
Deptment of Communications Equipment and Technologies
Technical University of Gabrovo
Gabrovo, Bulgaria
teddy.tedun@gmail.com

Georgi Georgiev
Department of Communications Equipment and Technologies
Technical University of Gabrovo
Gabrovo, Bulgaria
givanow@abv.bg

Abstract. **The article proposes an approach based on the concept of Artificial Intelligence for the categorization of urban areas of Internet content by corporate customers. The applicability of different neural apparatus was analyzed as well as three-layer Backpropagation Neural Networks (BPN) and four-layer Probabilistic Neural Networks (PNN) as the most suitable for the purpose of the study were selected. The synthesis of BPN architectures for Internet traffic identification was carried out according to a different number of computing units in the hidden layers with hyperbolic tangent sigmoid, log-sigmoid and linear transfer functions. The variations of a set of specific criteria were examined as Accuracy, Mean-Squared Error, Mean Absolute Error, Correlation coefficients, etc. The selection of PNNs against the defined quality indicators was based on a stepwise increase of the spread indicator of the Kernel functions in a Radial-Basis (RB) structural layer by analogy similar to that applied to BPNs. In the research processes, high levels of neural recognition indicators were established in processing with the Incoming flows of Internet Packages in an Accuracy of over 90.00%.**

Keywords: accuracy, backpropagation, probabilistic, traffic identification.

I. INTRODUCTION

The analysis of transmitted and serviced traffic in the global Internet network and ICT systems in relation to content consumption, optimization of transmission environment parameters and maintenance of QoS are important aspects of planning in modern communications. As tools for such procedures, a variety of Analytical Techniques, Methods, and Algorithms from "Artificial Intelligence" and "Machine Learning" concepts are applied [1, 2].

A substantial part of the research and development concerns the applicability of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks based on deep learning techniques. The main subjects of research are HTTP and SIP communications flows in connection with IoT traffic service [3]. CNN structures are often successfully used in connection with network traffic measurements for specific parameters such as throughput, packet sizes, times of receipt and release of user requests, repeatability in IP addressing addresses and ports, [4, 5] etc. RNN models can be useful in connection with the classification of encrypted traffic via FTP, HTTP, VoIP, XMPP and other protocol services in wireless sensor networks [6]. The combination of CNNs and Long Short-Term Memory (LSTM) when operating on VPN platforms allows for the effective identification of encrypted traffic content, including voice data, images and other formats [7]. In [8], an approach combining Deep Neural Networks (DNN) with k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM) solves various security monitoring and intrusion detection tasks regarding networks and network segments of IoT devices. Deep Learning Techniques, including Multi-Layer Perceptrons (MLPs), CNNs, RNNs, Generative Adversarial Networks (GAN), etc., are widely used in the classification of encrypted content in relation to well-known Google Applications [9]. Bayesian Neural Networks is a neural apparatus that meets the requirements of network operators in connection with the identification of anomalies in the transmission of traffic packages [10].

The article sets the task of categorizing the regions of consumption of Internet traffic by business companies in active time zones in an urban area with the help of

Print ISSN 1691-5402
Online ISSN 2256-070X
https://doi.org/10.17770/etr2023vol2.7265

Artificial Intelligence. The identification procedures are based on basic traffic parameters, respectively:

- Flows;
- Mean packet size in IPv4;
- Mean packet size in Ipv6;
- Mean packet transmission.

Systematized is a combined approach with the integration of Backpropagion Neural Networks and Probabilistic Neural Networks for analysis, evaluation and selection of models and categorization apparatus.

II. Meterial and Methods

The first phase of the research consists in the training and selection of Feed-Forward Neural Networks with Backpropagation learning processes. The training procedures are based on the Levenberg-Marquardt algorithm. The adequacy analysis of the selected neural apparatus on the classification task is based on the concept of different types of activation of the output layers of the BPN architectures and discrete code combinations to define the classification groups. In this case, neural models for identification of geographic zones of Internet content consumption were considered with:

- Linear activation type or "purelin";
- Hyperbolic tangent sigmoid transfer function or "tansig";
- Log-sigmoid activation type or "logsig".

The second phase of the research is aimed at assessing the quality of synthesized Backpropagation Neural Networks about performance analysis.

The final phase of the activities of the selection of neural apparatus and model for categorization of urban regions of Internet consumption is expressed in the assessment of the behavior of the Probabilistic Neural Networks. The architecture of the basic PNN model has been built by Input, Radial Basis, Competitive and Output Layers. By definition, the overall research framework for PNNs provides:

- constant state of neurons in the Radial Basis layer;
- lack of separation of phases of creation and training of neural models;
- change the "Spread" indicator on structural Kernel functions.

III. Results and Discussion

Traffic Zone Identification Using Feed-Forward Neural Networks with Backpropagation

The results of the quality assessment of the defined BPNs types for the criteria applied: 1) Accuracy, 2) Mean-Squared Error and 3) Mean Absolute Error are set out in Table 1 to Table 3.

Table 1 Results at BPN Selection Procedures in "purelin" Output Transfer Function

Hidden Neurons	Quality Indicators		
	Accuracy, %	***MSE***	***MAE***
3	91.3	0.0736	0.1653
4	95.7	0.0581	0.1149
5	95.7	0.0380	0.1276
6	95.7	0.0427	0.0577
7	95.7	0.0287	0.1050
8	100.0	0.0726	0.2123
9	95.7	0.0550	0.1345
10	95.7	0.0246	0.0963
11	95.7	0.0517	0.1783
12	87.0	0.1148	0.2150
13	95.7	0.0457	0.1218
14	100.0	0.0328	0.1327
15	100.0	0.0138	0.0970
16	100.0	0.0220	0.1096
17	91.3	0.0869	0.2292
18	95.7	0.0524	0.1297
19	91.3	0.0747	0.1982
20	95.7	0.0371	0.1412

In relation to the linear type of output activation, 95.7% accuracy was observed for the predominant of the created neural models. This fact applies to BPN structures with 4 to 7, 9-11, 13, 18 and 20 neurons in hidden layers. The mean squared error falls within the established range of 0.0138 for a network with 15 to 0.1148 about the categorization model containing 12 hidden neurons. While the MEA indicator registered a minimum of 0.0577 and a maximum of 0.2292 variations for the cases of BPNs at 6 and 17 neurons in structural hidden layer. Correct recognition of target samples from the input information set was achieved in neural architectures with a set amount of computational units in the hidden layer from 14 to 16. In view of the minimum MSE criteria requirement, it was identified as the best Backpropagation network with the presence of 15 hidden neurons.

Table 2 Results at Syntresis processes of BPNs in "tansig" Output Activation Type

Hidden Neurons	Quality Indicators		
	Accuracy, %	***MSE***	***MAE***
3	91.3	0.0904	0.1219
4	91.3	0.0912	0.1810
5	78.3	0.1371	0.2544
6	95.7	0.0321	0.0718
7	87.0	0.1308	0.1367
8	100.0	0.0040	0.0479
9	91.3	0.0532	0.0766
10	87.0	0.1149	0.1589
11	95.7	0.0468	0.1067
12	91.3	0.0762	0.1301

Hidden Neurons	Quality Indicators		
	Accuracy, %	*MSE*	*MAE*
13	91.3	0.0671	0.1727
14	95.7	0.0506	0.0973
15	95.7	0.0436	0.0481
16	95.7	0.0466	0.1168
17	95.7	0.0435	0.0439
18	82.6	0.1188	0.2146
19	95.7	0.0191	0.0557
20	91.3	0.0692	0.0827

With a view to the applied Hyperbolic tangent sigmoid output type of the activations, the following accuracies were found:

- low levels below 90.0%, respectively,82.6 % at 18, followed by 87.0 % for BPNs in 7 and 10 hidden neurons;
- 91.3% for neural architectures involving 3, 4, 9, 12, 13 and 20 neurons in the interlayer;
- equal to the previous level in terms of the number of neural structures, 95.7% for networks with fixed 6, 11, 14, 15, 16 and 17 hidden computing units;
- single accounted highest accuracy level of 100.0 % for a BPN model with 8 intermediate neurons.

In contrast to the previous BPN type studied, for the model with a correct classification of all samples, minimum readings of the basic MSE and MAE indicators, respectively 0.0040 and 0.0479 were established. Here, the errors recorded are significantly lower than their equivalents in architectures with a "purelin" output transfer function, determining the higher degree of suitability of the "tansing" activation type. The highest values of the indicated indicators MSE = 0.1371 and MAE = 0.2544 were found BPNs with 5 for the neural model with the lowest degree of recognition quality with Accuracy, which is equal to only 78.3%.

TABLE 3 RESULTS FOR BPNS SELECTION IN "LOGSIG" OUTPUT TRANSFER FUNCTION

Hidden Neurons	Quality Indicators		
	Accuracy, %	*MSE*	*MAE*
3	95.7	0.1431	0.2973
4	91.3	0.1672	0.3488
5	91.3	0.1684	0.2943
6	100.0	0.1358	0.2863
7	91.3	0.1558	0.2979
8	91.3	0.1590	0.3066
9	78.3	0.1825	0.3256
10	91.3	0.1663	0.3256
11	91.3	0.1471	0.3066
12	78.3	0.1929	0.3381
13	87.0	0.1968	0.3294
14	91.3	0.1660	0.3150
15	100.0	0.1326	0.2788
16	82.6	0.2097	0.3426
17	95.7	0.1439	0.2771
18	95.7	0.1478	0.2786
19	100.0	0.1293	0.2729
20	100.0	0.1337	0.2803

According to the used Log-sigmoid transfer function, a trend of similar levels of accuracy was found, but at times higher variations of MSE and MAE criteria. The Mean-Squared Error does not fall below 0.1200, and the Mean Absolute Error falls below 0.2700. Their minimum variations of 0.1293 and 0.2793 were reported for the neural structure with the highest rating of classification quality - accuracy 100.0%, at 19 hidden neurons. The highest error rates MSE = 0.2097 and MAE = 0.3488 were reached in BPNs in the composition, which were 16 and 4 neurons in the intermediate network layer. The analysed results give grounds for the considered neuron type to be categorized with the lowest degree of adequacy with regard to the assigned classification task.

Performance Analysis of BPN models for Internet Traffic Zone Categorization

The selected structural models for traffic identification in sequential setting of linear, hyperbolic tangent sigmoid and log-sigmoid neural activation of the output layers are given in Fig. 1.

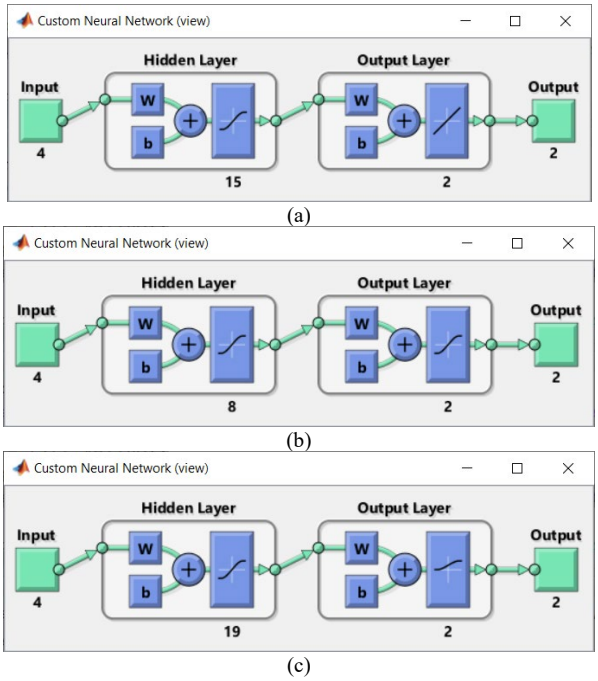
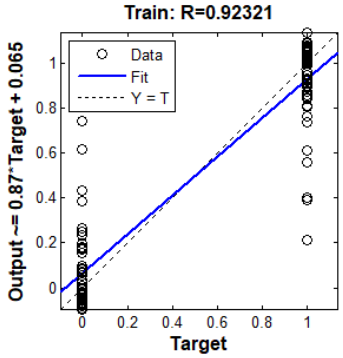
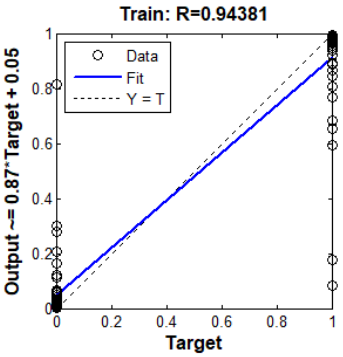


Fig. 1. Synthesized BPN architectures for Internet traffic zone categorization in (a) "purelin", (b) "tansig" and (c) "logsig" output activation types.

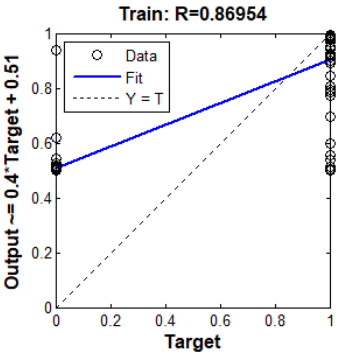
Regression diagnostic activities were carried out against the main network processes of training, validation and testing. Those procedures shall cover 70%, 15 % and 15 % of the volume of the input information sample. The linear regression dependencies for the state of network training generated for the selected BPNs are presented in Fig. 2. A good approximation of the theoretical and empirical regression lines can be seen against the "purelin" and "tansig" activation types. While the neuron model with a fixed Log-sigmoid transfer function (Fig. 2.(c)) shows a significant deviation, another significant sign of deterioration in the quality of the considered BPN is the underestimation of the correlation factor R below the threshold of 0.9000. Quantitative indicator levels R = 0.92321 for "purelin", R = 0.94381 at "tansig" and R = 0.86954 in "logsig" activations were reported, showing the best qualities of the second applied output activation.



(a)



(b)

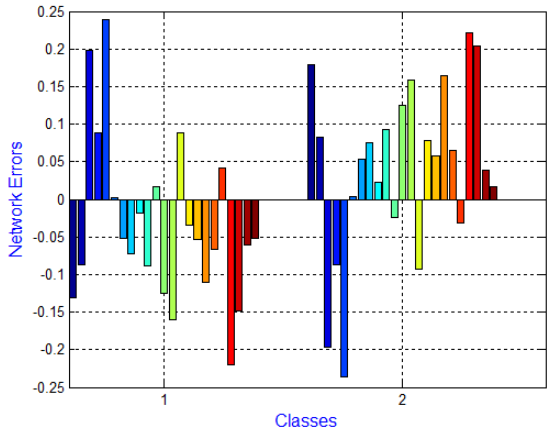


(c)

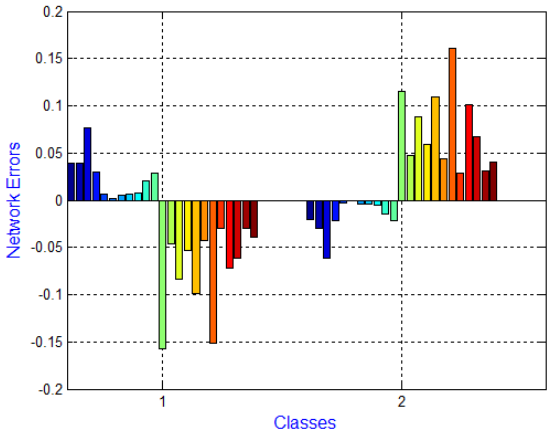
Fig. 2. Linear regression dependences for the learning process for selected BPNs for Internet traffic zone categorization in (a) "purelin", (b) "tansig" and (c) "logsig" output activation types.

TABLE 4 REGRESSION COEFFICIENTS ABOUT OUTPUTS OF SELECTED BPNS FOR TRAFFIC ZONE

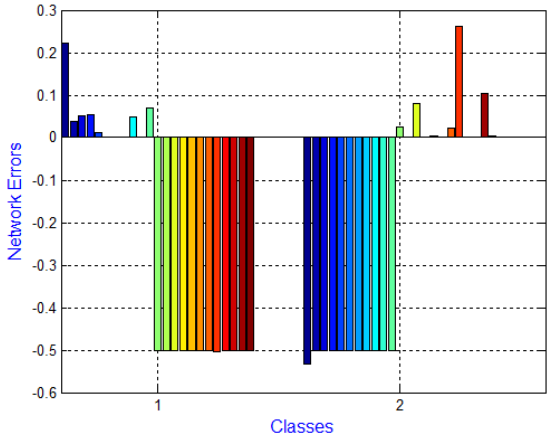
Network Output	Transfer Function		
	purelin	*tansig*	*logsig*
	R Indicator		
1 (Zone 1)	0.93475	0.93894	0.90516
2 (Zone 2)	0.93725	0.93927	0.9359



(a)



(b)



(c)

Fig. 3. Network error diagram for selected BPN models for Internet traffic zone categorization in (a) "purelin", (b) "tansig" and (c) "logsig" output transfer functions.

A similar trend regarding the lower suitability of the Log-sigmoid model is particularly confirmed for the first category zone of Internet consumption (Table 4). Here R reaches 0.90516 compared to levels around 0.9300 for linear and close to 0.9400 for hyperbolic tangent type. Compared to the second target class, minimal differences were registered, falling within the range of 0.9355 to 0.9330. The indicated data from the analysis of the relationship between the target and the output results from the use of neural models show an advantage of BPN in "tansig", followed by the one with "linear" and lastly of the network with fixed "logsig" output type.

The established qualities of the models are confirmed by the diagrams of network errors (the differences between the desired and the results calculated with the attached BPN) of Fig. 3. In the first two activation types of neural outputs, satisfactory levels below the 0.5000 limit were achieved. Ranges "-0.2358 to 0.2392" and "-0.1575 to 0.1614" have been established in relation to the use of linear and hyperbolic tangent sigmoid transfer functions. In the course of the study, levels were significantly higher than those recorded in the previous two types of models below the negative threshold of "-0.5000" with a maximum peak of "-0.5339" for a large part of the data from the composition of the test subset. The maximum observed upper threshold of network error at BPN with Log-sigmoid is "0.2627", commensurate with the "purelin" activation.

Traffic Zone Identification with Probabilistic Neural Networks

The main PNN model in connection with geraral investigation procedures is given in Fig. 4. Regarding the selection processes, an assessment of the MSE and MAE criteria was performed with a step increase of the Spread in the range from 0.05 to 0.90.

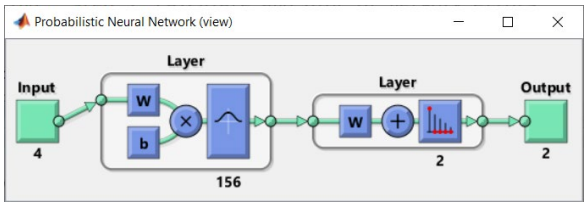


Fig. 4. PNN model for identification of Internet traffic zones.

In the initial range to a level around "0.20", the registered error levels were found to be in the range of "10-5" degree (significantly higher compared to those registered with BPN architectures). A gradual increase in the Mean-Squared Error and Mean Absolute Error indicators was observed, reaching approximate levels of "0.0060" and "0.0500" at the end of the Spread range. As a result of the overall analysis, a probabilistic neural model was selected when setting Spread = 0.10 with respect to the Radial Basis structural layer.

Figure 5 presents the distribution of the samples forming the input information set in relation to the classification procedure for each output group. There is a correct determination of the group affiliation of the test data - 100.0% levels have been obtained for Sensitivity, Precision and Accuracy indexes.

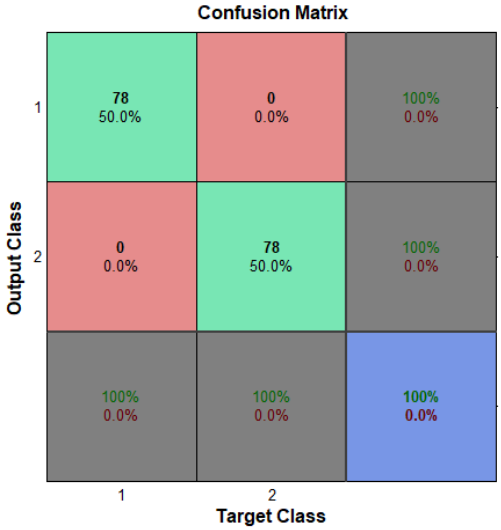


Fig. 5. Confusion matrix about PNN architectures in Internet traffic zones identification.

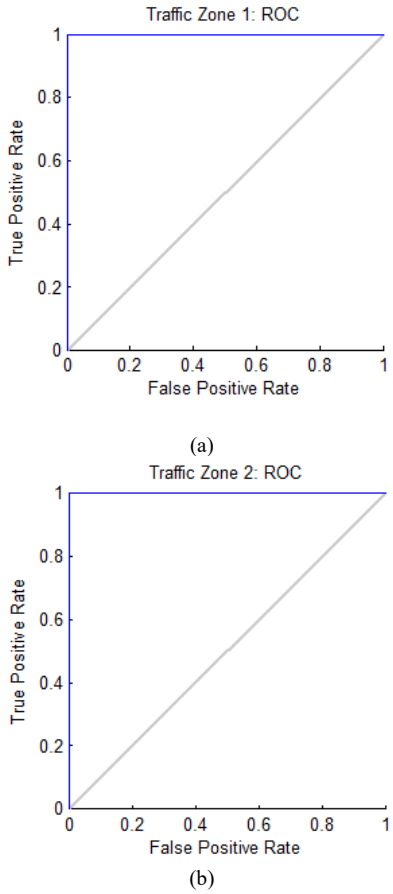


Fig. 6. ROC dependencies for PNN outputs in Internet traffic zone categorization – (a) class № 1 and (b) class № 2.

IV. CONCLUSION

In view of the aggregated results, the proposed approach to categorizing Internet traffic using BPNs and PNNs shows good applicability. In this case, an advantage of probabilities over neural structures with reverse propagation of the error has been established. Synthesized models should be embedded in ICT traffic monitoring units supporting the activities of system administrators. In the next phase of the research, activities are set for adapting models and optimization procedures in relation to predictive analysis in traffic distribution planning.

Acknowledgments

This research was funded and supported by Project "Development of ICT-based systems for investigation and monitoring of traffic and user access by Artificial Intelligence" in Technical University of Gabrovo, Bulgaria.

REFERENCES

[1] D. Petrova, Analysis of SMEs in Bulgaria – Assessment of their innovation activities: International Scientific and Practical Conference, July 20-22, 2013, Rezekne, Latvia.

[2] D. Petrova, Innovativeand sustainable industry in Bulgaria – prospects and challenges: International Scientific and Practical Conference „Environment, Technology and Resources", June 20-22, 2019, Rezekne, Latvia.

[3] M. L. Martin, B. Carro, A. Esguevillas, and J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for internet of things," IEEE Access, vol. 5, pp. 18042-18050, Sep. 2017.

[4] G. Freine, "Deep learning for the analysis of network traffic measurements," M.S. thesis, Universidad de la Republica, Montevideo, ON, Uruguay, 2019.

[5] G. Mihaylov, T. Iliev, I. stoyanov, and E. Ivanova, An approach for point-to-point link within mobile network coverage: International Scientific Conference on Communications, Information, Electronic and Energy Systems,November 25-27, 2022, Ruse, Bulgaria.

[6] W. Aitken and D. Brown, "Application traffic classification using neural networks," Defence Research and Development Canada, Canada, Tech. Rep. DRDC-RDDC-2022-R052, 2022.

[7] X. Hu, Ch. Gu, and F. Wei, " A Network Combining CNN and LSTM for Internet Encrypted Traffic Classification," Hindawi, Security and Communication Networks, vol. 2021, pp. 1-15, June 2021.

[8] M. B. Umair, Z. Iqbal, M. Bilal, J. Nebhen, T. Almohamed, and R. Mehmood, "An Efficient Internet Traffic Classification System Using Deep Learning for IoT," Computers, Materials & Continua, vol. 71, pp. 407-422, Nov. 2022.

[9] Sh. Rezaei and X. Iu, "Deep Learning for Encrypted Traffic Classification: An Overview," IEEE Communications Magazine, vol. 57, pp. 76-81, May 2019.

[10] P. Michael, E. Valla, and N. Neggatu, " Network traffic classification via neural networks," University of Cambridge, United Kingdom, Tech. Rep. UCAM-CL-TR-912, Sep. 2017.